Relevant distance between two different instances of the same potential energy in protein folding

José L. Alonso [a,b] Pablo Echenique [a,b,*]

[a] *Instituto de Biocomputación y Física de los Sistemas Complejos (BIFI), Edificio Cervantes, Corona de Aragón 42, 50009, Zaragoza, Spain.*

[b] *Departamento de Física Teórica, Facultad de Ciencias, Universidad de Zaragoza, Pedro Cerbuna 12, 50009, Zaragoza, Spain.*

Abstract

In the context of complex systems and, particularly, of protein folding, a physically meaningful distance is defined which allows to make useful statistical statements about the way in which energy differences are modified when two different instances of the same potential-energy function are used. When the two instances arise from the fact that different algorithms or different approximations are used, the distance herein defined may be used to evaluate the relative accuracy of the two methods. When the difference is due to a change in the free parameters of which the potential depends on, the distance can be used to quantify, in each region of parameter space, the robustness of the modeling to such a change and this, in turn, may be used to assess the significance of a parameters' fit. Both cases are illustrated with a practical example: the study of the Poisson–based solvation energy in the Trp–Cage protein (PDB code 1L2Y).

Key words: protein folding, Poisson equation, Poisson-Boltzmann equation, potential accuracy, solvation energy, distance criterium
PACS: 87.15.-v, 87.15.Aa, 87.15.Cc, 87.14.Ee, 41.20.Cv

1 Introduction

The most fundamental way to account for the behaviour of a physical system is through its energy function $\mathcal{H}(\vec{q}, \vec{p})$, which depends on the coordinates $\vec{q}$

---

* Correspondent author. *Email address:* `pnique@unizar.es`

and the momenta $\vec{p}$ of all the particles. In normal situations, this function can be expressed as the sum of the kinetic energy $\mathcal{K}(\vec{q}, \vec{p})$ [1] and the potential energy $\mathcal{V}(\vec{q})$. Since the former is of general form for any type of system and, normally, it does not affect the equilibrium properties, the latter is enough for a complete characterization of the problem.

Under most real circumstances, the exact form of $\mathcal{V}(\vec{q})$ is unknown and one is forced to seek an approximation $\mathcal{V}^{\mathrm{app}}(\vec{q})$. This may be done, for physical systems that are significantly complex, by assuming that the relevant interactions included in $\mathcal{V}(\vec{q})$ can be formally factorized [1]. Then, an approximated function $\mathcal{V}_i^{\mathrm{app}}(\vec{q})$ is devised according to heuristic and semiempirical reasons to account for each of the original parts $\mathcal{V}_i(\vec{q})$:

$$\mathcal{V}(\vec{q}) = \sum_{i=1}^{m} \mathcal{V}_i(\vec{q}) \simeq \sum_{i=1}^{m} \mathcal{V}_i^{\mathrm{app}}(\vec{q}) \,. \tag{1}$$

For example, in the study of proteins [2,3], which are a very relevant case of complex systems, some of the terms in which the total potential-energy function is traditionally factorized are the hydrogen-bonds energy, the Van der Waals interaction, the excluded-volume repulsion, the Coulomb energy and the solvation energy. This last interaction, which is one of the most challenging terms of $\mathcal{V}(\vec{q})$ to model, is customarily further split into electrostatic and non-electrostatic parts [4,5]. It is the former which is studied in section 3 to illustrate the application of the concepts herein discussed.

Returning to the general case, let us assume that a particular energy term $\mathcal{V}_j(\vec{q})$ in eq. 1 and its approximated counterpart $\mathcal{V}_j^{\mathrm{app}}(\vec{q})$ correspond to the part of $\mathcal{V}(\vec{q})$ that is going to be studied or modeled. Let us denote that term $\mathcal{V}_j(\vec{q})$ by $V(\vec{q})$ and, correspondly, $\mathcal{V}_j^{\mathrm{app}}(\vec{q})$ by $V^{\mathrm{app}}(\vec{q})$ in the forecoming reasoning. Clearly, if the approximated function $V^{\mathrm{app}}(\vec{q})$ is *too distant* from the original $V(\vec{q})$, it will be useless, as this difference will certainly propagate to the total potential energy. Therefore, one must precisely define and calculate this *distance*, depending on the type of system which is the object of the study and on the particular aspects that are going to be investigated. The situation is further complicated when subsequent approximations to $V^{\mathrm{app}}(\vec{q})$ are done, usually with the aim of lessening the numerical complexity and rendering the simulations feasible.

This yet-undefined distance between potential-energy functions may also be useful in another situation which is often found in the study of complex systems: parameter fitting. Any reasonable functional form of a certain term $V(\vec{q})$ of the total potential energy (or of its approximation $V^{\mathrm{app}}(\vec{q})$) is a simplified

---

[1] The kinetic energy $\mathcal{K}$ depends, in general, on the positions and the momenta. However, if cartesian coordinates are used, the dependence on positions vanishes.

model of physical reality and it contains a number of free parameters. These parameters, which, in most of the cases, are not physically observable, must be fit against experimental or more *ab initio* results prior to using the function for practical purposes. For example, in the continuum solvation models based on the solution of the Poisson equation [5,6,7,8], typical free parameters are the dielectric constants and the position of the surface that separates the outer high–dielectric medium from the inner low–dielectric one. Although they are customarily assigned *standard* values (such as $\kappa_P = 1$ for the dielectric constant of the protein, $\kappa_W = 80$ for the dielectric constant of the aqueous medium[2], and the Molecular Surface (MS), defined by Connolly [9], for the surface of separation), we believe that they must be fit in order to render calculations more accurate. Certainly, for other potential-energy functions, such as the ones found in force fields like CHARMM [10,11], the fit of the free parameters is common practice.

In order for any fit to yield statistically significant values of the parameters, the particular region of the parameter space in which the final result lies must have the property of *robustness*, i.e., it must occur that, if the found set of parameters' values is slightly changed, then, the relevant characteristics of the potential-energy function which depends on them are also approximately kept unmodified. If this were not the case, a new fit, performed using a different set of experimental (or more *ab initio*) points, could produce a very *distant* potential. This last scenario is, clearly, undesirable. Therefore, it may be useful to evaluate the robustness of the, *a priori*, reasonable regions of parameter space for the potential energy function that is to be used. To accomplish this, one must again define a relevant distance between two *instances* of the same potential with different values of the parameters.

In section 2, a meaningful distance that can be used in the two situations aforementioned is defined and justified. In section 3, within the context of the protein folding problem and as an example of the first application discussed, this distance is measured between instances of the Poisson–based solvation energy arising from the choice of different grid sizes in the finite-differences algorithm with which it is calculated. To illustrate the second possible use of the distance, the robustness of the Poisson energies to changes in some of its free parameters, holding the grid size fixed, is quantified. This analysis is necessary to assess the significance of the parameters' values determined through a fit. Finally, section 4 is devoted to the conclusions.

---

[2]   In the field of molecular simulations, κ usually denotes the Debye-Hückel parameter, which quantifies the ionic strength in the aqueous medium; ϵ is customarily used to represent the dielectric constant. However, in this work, the usual convention in physics, by which κ stands for the dielectric constant and ϵ stands for the dielectric permittivity, is preferred. Since the ionic strength is zero in all calculations, this choice should not be misleading.

2 Distance criterium

Let $V(\vec{q})$ be a particular term of the potential energy of a system. The numerical value of this physical quantity for each conformation $\vec{q}$ depends on two conceptually different inputs: on one hand, the algorithm or approximation used to compute it, denoted by $\mathcal{A}$; on the other hand, the values of the free parameters $\vec{P}$. Changes in these inputs produce different *instances* of the physical quantity $V(\vec{q})$, which we denote by subscripting V. For example, if the algorithm is held constant and two different set of parameters $\vec{P}_1$ and $\vec{P}_2$ are used, our notation made explicit would read as in the following equation (an analogous definition may be made if the algorithm is held constant and the parameters are varied).

$$V_1(\vec{q}) := V(\mathcal{A}, \vec{P}_1, \vec{q}) \qquad \text{and} \qquad V_2(\vec{q}) := V(\mathcal{A}, \vec{P}_2, \vec{q}) \,. \tag{2}$$

Now, a useful and physically meaningful definition of a distance $d(V_1, V_2)$ is sought between a pair of instances, such as the ones in the previous equation, of the same potential-energy function.

In some cases traditionally studied in physics, the dependence of V on the parameters is simple enough to allow a closed functional dependence $V_2(V_1)$ to be found [3]. However, in the study of complex systems, such as proteins, this dependence is often much more complicated, due to the high dimensionality of the conformational space and to the fact that the energy landscape lacks any evident symmetry. The set $\mathcal{C}(V_1)$ of the conformations with a particular value of the potential energy V_1 typically spans large regions of the phase space containing structurally different conformations. When an approximation is performed or the algorithm is changed, from $\mathcal{A}_1$ to $\mathcal{A}_2$, or the free parameters are shifted, from $\vec{P}_1$ to $\vec{P}_2$, each conformation $\vec{q}$ in $\mathcal{C}(V_1)$ is affected in a different way and its new energy $V_2(\vec{q})$ is modified in a manner that does not depend trivially on the particular region of the phase space which the conformation $\vec{q}$ belongs to. Therefore, a simple functional relation $V_2(V_1)$ is no longer possible to be found: for each value of V_1, there corresponds now a whole distribution of values of V_2 associated with conformations which share the same value of V_1 but which are far apart in the conformational space. Moreover, the projection of this high–dimensional $\vec{q}$–space into the 1–dimensional V_1–space makes V_2 look as a random variable for each particular value of V_1 (see fig. 1). We then

---

[3] For example, if the mass of a harmonic oscillator is changed from m_1 to m_2, the potential energy functions will satisfy the linear relation $V_2(\vec{q}) = (m_1/m_2)V_1(\vec{q})$ for all the conformations of the system; if the the atomic charges are rescaled by a factor α (being actually αQ_i) and α is changed from α_1 to α_2, the free energies of solvation calculated via the Poisson equation will, in turn, satisfy the linear relation $V_2(\vec{q}) = (\alpha_1/\alpha_2)^2 V_1(\vec{q})$, etc.
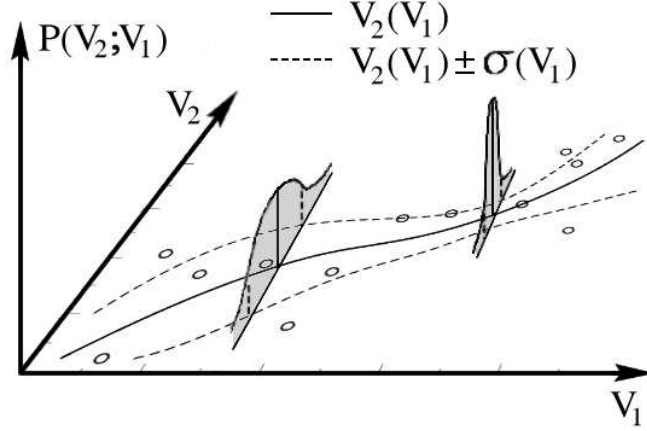
Fig. 1. *Illustration* of the functions defined in eq. 3. Each point corresponds to a single conformation $\vec{q}_i$ of the system, being $V_1(\vec{q}_i)$ its x–coordinate and $V_2(\vec{q}_i)$ its y–coordinate. In the z–axis, the probability density of *measuring* a particular value of V_2 for each value of V_1. Two normal distributions representing this quantity are shown at arbitrary positions in the x–axis. The continuum line represents the mean $V_2(V_1)$ of the values $V_2(\vec{q})$ with $\vec{q} \in \mathcal{C}(V_1)$ as a function of V_1. The broken lines enclose the region where there is the largest probability to find a conformation if a single *numerical experiment* is performed (around 68% if the distribution of V_2 is assumed to be normal for each V_1).

define two real functions, $V_2(V_1)$ and $\sigma(V_1)$, which correspond to the mean and to the standard deviation, respectively, of this random variable as a function of V_1 (where the average $\langle \cdot \rangle$ is defined to be taken over the conformations $\vec{q} \in \mathcal{C}(V_1)$):

$$V_2(V_1) := \langle V_2(\vec{q}) \rangle \qquad \text{and} \qquad \sigma(V_1) := \sqrt{\langle (V_2(\vec{q}) - \langle V_2(\vec{q}) \rangle)^2 \rangle} \,. \tag{3}$$

It can be proved, from their definition, that these two functions are continuous irrespectively of the particular characteristics of $V_1(\vec{q})$ and $V_2(\vec{q})$ (given that both of them are *smooth* functions of the conformation). It may also be shown that, under assumptions which are typically fulfilled in real cases, $V_2(V_1)$ and $\sigma(V_1)$ are also differentiable. Thus, when restricted to a finite interval of V_1, the linear dependence given by the following equation may hold approximately:

$$V_2(V_1) \simeq bV_1 + a := b(V_1 + V_1^0) \,. \tag{4}$$

In fact, for the aforementioned cases in which the dependence of the potential energy on the parameters is simple enough (see footnote 3), eq. 4 holds exactly and, being able to describe $V_2(V_1)$ by a closed analytical formula, b and a can be exactly computed[4]. However, in a general situation, the functions

---

[4] One must be careful about the notation. Although the function $V_2(V_1)$ has been formally defined in eq. 3 as an average, in the case of the simple systems in foot-

defined in eq. 3 are impossible to be calculated analytically and so are the parameters b and a in the linear approximation of $V_2(V_1)$. In such a case, one may at most have a finite collection of n conformations $\{\vec{q}_i\}_{i=1}^n$ and the respective values of $V_1(\vec{q}_i)$ and $V_2(\vec{q}_i)$ for each one of them. These data, for a particular i, should be thought as a single *numerical experiment* in the already suggested sense that, if one regards $V_1(\vec{q}_i)$ as an independent variable, the *outcome* of the dependent variable $V_2(\vec{q}_i)$ is basically unpredictable and V_2 may be regarded as a random variable parametrically dependent on V_1 (see fig. 1). From this finite knowledge about the system, one may statistically estimate the values of b and a. If the standard deviation $\sigma(V_1)$ is a constant (i.e., it does not depend on V_1), which, for the particular system studied in this work, has been found to be approximately true, then, the least-squares maximum-likelihood method [12,13] yields the best estimates for b and a[5] under very general conditions, such as independence of the random variables and normal distribution. Moreover, it may be shown [12,13] that the best estimate for the standard deviation σ (see footnote 5) is given by the following expresion:

$$\sigma \simeq \sqrt{\frac{\sum_{i=1}^n [V_2(\vec{q}_i) - (bV_1(\vec{q}_i) + a)]^2}{n - 2}} \; . \tag{5}$$

This quantity σ may be regarded as a *random error* that arises in the transit from V_1 to V_2. In the same sense, a may be regarded as a *systematic error* and, since it is equivalent to an energy reference, its actual value is not relevant for the forecoming discussion. The slope b and σ are the two quantities involved in the definition of the distance $d(V_1, V_2)$, which is the central concept introduced in this paper (see eq. 8 below). In order to render this definition meaningful, we are going to evaluate how *energy differences* are modified, when going from V_1 to V_2, as a function of b and σ. This differences are the relevant physical quantities if one's aim is to study the conformational behaviour of a system.

Under the approximations of linear $V_2(V_1)$ dependence and constancy of $\sigma(V_1)$, fig. 1 becomes the left part of fig. 2. In the right side of the same figure, one of the *worst* cases (i.e., one case for which $d(V_1, V_2)$ is one of the largest) studied in section 3 is depicted to show that the hypothesis are approximately fulfilled for the particular protein system investigated there.

Now, let us focus on two arbitrary conformations of the system (see the left side of fig. 2). For them, the V_2–energy difference ΔV_2 is a random variable

---

note 3, a function $V_2(V_1)$ may be found algebraically. This is, however, consistent. If one thinks that, in these systems, the spread $\sigma(V_1)$ is zero, the relation found algebraically and the one that stems from the average are identical.

[5] The same letters are used for the ideal parameters b, a, and σ and for their least-squares best estimates, since the only knowledge that one may have about the former comes from the calculation of the latter.
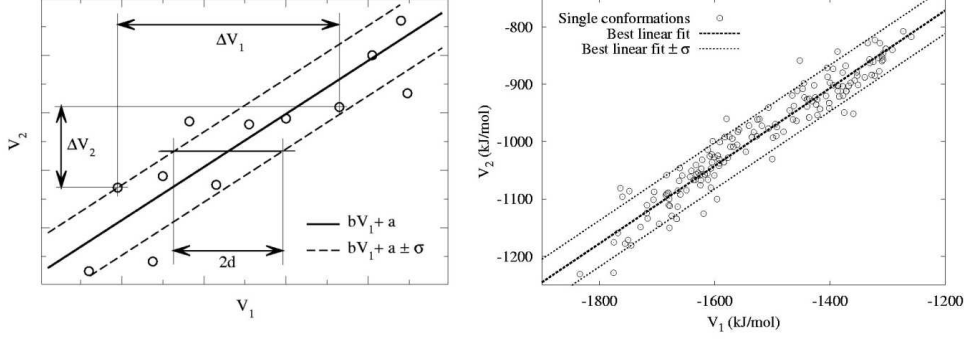
Fig. 2. The graphic on the left is an *illustration* of the distance criterium. The best linear fit is depicted by a continuum line. The broken lines correspond to the estimated standard deviation of the points in the V_2 direction. The V_1- and V_2–energy differences, ΔV_1 and ΔV_2, between two particular conformations, as well as the distance d for $\alpha = 1$ (see eq. 8) are also shown. The graphic on the right is a particular example taken from the system studied in section 3. The difference between the V_1 and V_2 instances is a small modification of the surface of separation between dielectrics. 150 conformations of the investigated protein are shown.

parametrically dependent on the V_1–energy difference ΔV_1 (which is assumed to be regular number, i.e., a random variable with zero variance [6]). The probability density of this random variable is found by assuming that V_2 is normally distributed with mean $V_2(V_1)$ and standard deviation σ. Then, the distribution of ΔV_2, for each ΔV_1, is normal with mean $b\Delta V_1$ and standard deviation $\sqrt{2}\sigma$ [7] :

$$\mathcal{P}(\Delta V_2; \Delta V_1) = \frac{1}{\sqrt{2\pi}(\sqrt{2}\sigma)} \exp\left[-\frac{(\Delta V_2 - b\Delta V_1)^2}{2(\sqrt{2}\sigma)^2}\right] . \tag{6}$$

If the random errors were negligible (as in the systems discussed in footnote 3) and one wanted to calculate the value of ΔV_1 from the knowledge of ΔV_2, the identity $\Delta V_1 = \Delta V_2/b$ would have to be used. When there are significant random errors, the situation is equivalent except for the fact that there is a probabilistic indetermination, i.e., from the *measured* value of ΔV_2, ΔV_1 can be no longer inferred. It follows directly from eq. 6 that, the quantity $\Delta V_2/b$ is a random variable normally distributed with mean ΔV_1 and standard deviation $\sqrt{2}\sigma/|b|$ (see footnote 7):

$$\mathcal{P}(\Delta V_2/b; \Delta V_1) = \frac{1}{\sqrt{2\pi}(\sqrt{2}\sigma/|b|)} \exp\left[-\frac{(\Delta V_2/b - \Delta V_1)^2}{2(\sqrt{2}\sigma/b)^2}\right] . \tag{7}$$

---

[6] See the discussion near the end of the section for a more detailed analysis of the implications of this assumption.

[7] If x, y and z are random variables and the relation $z = Ax + By$ holds, then $\langle z \rangle = A\langle x \rangle + B\langle y \rangle$. If x and y are independent, one also has that $\sigma_z^2 = A\sigma_x^2 + B\sigma_y^2$ [12,13].

Now, let us define the distance $d(V_1, V_2)$ between two instances of the same potential energy as follows[8]:

$$d(V_1, V_2) := \alpha \frac{\sigma}{|b|} = \frac{\alpha}{|b|} \sqrt{\frac{\sum_{i=1}^{n}[V_2(\vec{q}_i) - (bV_1(\vec{q}_i) + a)]^2}{n - 2}} \ . \tag{8}$$

Where b and a are those calculated from a least-squares fit of the values of $V_2(\vec{q}_i)$ against $V_1(\vec{q}_i)$ and α is a positive proportionality factor yet to be fixed (see point 3 below).

This definition encodes some intuitions that one may have about the *loss of information* involved in the transit from V_1 to V_2. Let us remark some important properties which illustrate this fact:

(1) If the slope b is nonzero and the random error σ is zero, $d(V_1, V_2) = 0$ and there is no loss of information. An example of this situation is given by the simple systems in footnote 3: clearly, no loss of information may be involved in changing the mass of an harmonic oscillator.

(2) If the random error σ is different from zero and the slope b goes to zero, $d(V_1, V_2) \to \infty$ and the loss of information is complete. One may picture this situation by making the best-fit line in fig. 2 horizontal. In such a case, when two numerical experiments are performed, the probability distribution of measuring ΔV_2 does not depend on ΔV_1 (it is normal with zero mean) and the information about ΔV_1 is impossible to recover.

(3) For intermediate cases in which both b and σ are nonzero, all the information about $\Delta V_2/b$ is found in its probability-density function (see eq. 7) and many probabilistic statements may be made. For example, it would be desirable that the sign of $\Delta V_2/b$ had a high probability of being equal to the sign of ΔV_1. This would typically keep the correct energy ordering of the conformations when going from V_1 to V_2. Making the variable change $x = \Delta V_2/b - \Delta V_1$ in eq. 7, one finds that this probability is given by the following equation (assuming, without loss of generality, that $\Delta V_1 > 0$).

$$\mathcal{P}_{\text{ordering}} = \int_{-\Delta V_1}^{\infty} \frac{1}{\sqrt{2\pi}(\sqrt{2}d/\alpha)} \exp\left[-\frac{x^2}{2(\sqrt{2}d/\alpha)^2}\right] \mathrm{d}x \ . \tag{9}$$

For a same value of ΔV_1, this probability decreases with d; if d is held constant, it increases with ΔV_1 (the minimum being $\mathcal{P}_{\text{ordering}} = 1/2$, either for $\Delta V_1 \to 0$ or for $d \to \infty$). If d is small, the probability of the

---

[8] Although a negative value of b may look physically perverse, there is no theoretical drawback about it and the possibility is allowed. In fact, if the random errors were small and b was not very small in absolute value, the loss of information (see the forecoming discussion) in going from V_1 to V_2 would be small.

ordering being mantained is large, even for pairs of conformations that are close in V_1–energy. For example, if one takes $\alpha = 1$, it happens that, if $\Delta V_1 > d$, then, $\mathcal{P}_{\text{ordering}} > 76\%$; if $\Delta V_1 > 2d$, then, $\mathcal{P}_{\text{ordering}} > 92\%$, etc. Any other choice of α would only yield different numerical values for this bounds on $\mathcal{P}_{\text{ordering}}$; the qualitative facts would be preserved. However, since this values are natural (the normal distribution varies rapidly and one would easily get very large or very small probabilities), the choice $\alpha = 1$ is considered to be the most convenient herein and it is the one to be used in section 3.

(4) The properties stated in the previous point are direct consequences of the more general fact that, as d decreases, so does the standard deviation of the random variable $\Delta V_2/b$ (which equals $\sqrt{2}d/\alpha$) and the distribution becomes sharper around the average ΔV_1 (see eq. 7). That is, the smaller the value of d, the larger the probability of $\Delta V_2/b$ being close to the *perfect* value ΔV_1.

This measure of the distance between two instances of a potential energy presents some advantages. On one hand, if the approximations on which it is based (normal distribution of V_2 for each V_1, linear $V_2(V_1)$ dependence, constancy of $\sigma(V_1)$ and zero variance in the *measures* of V_1) are reasonable, the statistical statements derived from a particular d value are meaningful and precise. On the other hand, this statements refer, like the ones in the points discussed above, to the whole conformational space. However, we would like to stress that, in this work, we are not going to give any criterium to decide whether a particular value of d is sufficiently small for an approximation $V_1 \rightarrow V_2$ to be valid or for the system to be robust to changes in the free parameters. Such a decision must be taken depending on the particularities of the system studied (which are encoded in the total potential energy function $\mathcal{V}(\vec{q})$) and on the questions sought to be answered. Our definition of $d(V_1, V_2)$, being based in characteristics shared by many complex systems, is of general application. For example, we are not going to establish any limit on the accuracy required for a potential energy function to successfully predict the folding of proteins [2,3]. We consider this question a difficult theoretical problem and we believe that it may be possible *a priori* that some special features of the energy landscapes of proteins (such as funnel–like shape) are the main responsible of the high efficiency and cooperativity of the folding process [2,3]. If this were the case, a different procedure for measuring the distance between potential energy functions could be devised [14,15,16], as any approximation which does not significantly alter these special features would be valid even if the value of d is very large. However, for the sake of simplicity, it will be assumed, herein and in section 3 that a transit $V_1 \rightarrow V_2$ between instances of the same potential whose $d(V_1, V_2)$ value is *of the order* of the thermal fluctuations RT

is acceptable [9].

The last point that we must remark in this section is that the distance introduced in this paper does not satisfy all the properties that the class of mathematical objects usually referred to as *distances* do satisfy. For example, the equivalence $\mathcal{D}(x,y) = 0 \Leftrightarrow x = y$ becomes, for the distance in eq. 8, an implication in only one direction, i.e., while it is true that $V_1 = V_2 \Rightarrow d(V_1, V_2) = 0$, the reciprocal is false in general, since, for example, if $V_2 = BV_1 + A$, with B and A non-zero constants, then $d(V_1, V_2) = 0$ when, clearly, the two instances are not equal. Another property of the mathematical distances that is not fulfilled by d is the one of symmetry, i.e., that $\mathcal{D}(x,y) = \mathcal{D}(y,x)$. If we denote all the quantities calculated when going from V_1 to V_2 by subscripting them with the label 12 and the ones corresponding to the opposite process with 21, we have that:

$$d(V_1, V_2) = \alpha \sqrt{\frac{\sum_{i=1}^{n}[V_2(\vec{q}_i)/b_{12} - (V_1(\vec{q}_i) + a_{12}/b_{12})]^2}{n-2}} \; ,$$
$$d(V_2, V_1) = \alpha \sqrt{\frac{\sum_{i=1}^{n}[V_1(\vec{q}_i)/b_{21} - (V_2(\vec{q}_i) + a_{21}/b_{21})]^2}{n-2}} \; . \tag{10}$$

Therefore, if the equality $d(V_1, V_2) = d(V_2, V_1)$ is to be hold for every set of conformations $\{\vec{q}_i\}_{i=1}^n$, one must require that $b_{12} = b_{21} = 1$ and $a_{12} = -a_{21}$. These two relations impose complicated conditions over the values of $\{V_1(\vec{q}_i)\}_{i=1}^n$ and $\{V_2(\vec{q}_i)\}_{i=1}^n$ which are not generally fulfilled. The origin of this lack of symmetry is completely consistent with the assumptions made about the random character of the two instances of the potential energy to be compared, namely, the hypothesis of zero variance of V_1, which places the two potentials on a different footing. A more general distance may be defined (J.L. Alonso and P. Echenique, in preparation) that takes into account a possible indetermination in the *measure* of V_1 and that places the two potentials on the same footing. However, some remarks must be made about this. In the first place, this asymmetry in the role played by each of the potential-energy instances is totally justified in the cases for which the situation intended to be modeled is actually asymmetric; for example, if one's aim is to calculate the distance between a potential V_1 and an approximation V_2, where V_1 may be considered either *ex-*

---

[9] In most computational methods and theoretical descriptions of a system in contact with a thermal reservoir RT is a relevant energy (RT is preferred to $k_B T$ since per mole energy units are used in this article) and the results will be presented in units of RT. It appears in the thermodynamical equilibrium Boltzmann distribution, in which the probability p_i of a conformation $\vec{q}_i$ is proportional to $\exp(-V(\vec{q}_i)/RT)$, it also determines the transition probability $\min[1, \exp(-(V(\vec{q}_{i+1}) - V(\vec{q}_i))/RT)]$ in the Metropolis Monte Carlo scheme and it is the spread of the stochastic term in the Langevin equation [17].

act (e.g., quantum mechanical) or much more accurate than V_2. In such a case, the assumption of zero variance for V_1 and the difference in the roles played by both potentials is intrinsic to the situation studied. As a second remark, it must be stressed that the distance herein defined was never intended to be a mathematical distance, although some of the properties demanded to these objects are satisfactory and fairly intuitive. The meaning of the distance d is encoded in the statistical statements derived from its value and the name *distance* must be used in a more relaxed manner than the one traditionally found in mathematics. Finally, it must be pointed out that there is a situation in which the symmetry of the distance defined in this work holds, namely, the situation in which $b := b_{12} = 1$ and $n \to \infty$. When the number of conformations n is very large, the statistical estimators b and a of the slope and the y–intercept of the linear relation between V_1 and V_2 tend to the ideal values (see footnote 5) and, in these conditions, the remaining requirements needed to satisfy symmetry are also fulfilled, i.e., one has that $b_{21} = 1$, $a_{12} = -a_{21}$ and, consequently, $d(V_1, V_2) = d(V_2, V_1)$ [10]. This fact is relevant since there are many situations typically found in which $b := b_{12} \simeq 1$, namely, those in which V_1 and V_2 are *proximate*. This is the case if one wants to assess the robustness of a potential-energy function (a slight change in the parameters does not lead to a completely different energy) or if the approximation performed is small. The two applications of the distance d to a particular potential in sec. 3 are carried out in cases for which this *proximity* is achieved and the symmetry expected has been checked numerically.

3 Application

Most of the finely tuned biomolecular events occur in a complex environment of unique characteristics: liquid water. Therefore, if one aims to correctly describe the crucial processes associated with proteins, DNA and RNA in living beings, a sufficiently accurate modeling of water–water and water–solute interactions must be implemented. However, accuracy is not the only criterium to be followed when designing a solvation model. Numerical complexity of the methods must be also taken into account, as computational power is always a limiting resource. A compromise between these two competing factors must be reached and precision may be traded for velocity, even if the understanding of the problem was complete and a great accuracy could be achieved.

Particularly, in the study of the protein folding problem, the search for the native state takes place in an astronomically large conformational space, as

---

[10] Note that, if one has $b := b_{12} = 1$ and $n \to \infty$, the implication $d(V_1, V_2) = 0 \Rightarrow V_1 = V_2 + A$ also hold. Since V_1 and V_2 are physical energies defined up to a reference, this may be considered the reciprocal of $V_1 = V_2 \Rightarrow d(V_1, V_2) = 0$.
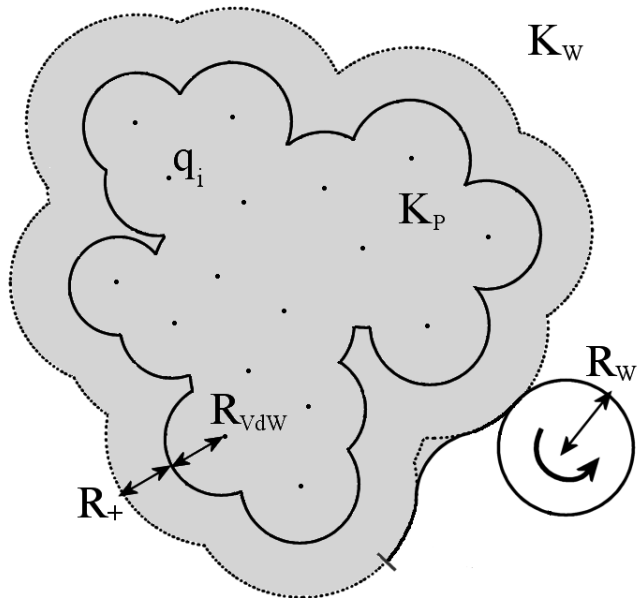
Fig. 3. Dielectric model of the protein. Atomic charges q_i are punctual and located at nuclei. Space is divided in two disjoint regions: the macromolecule (in gray), with low dielectric constant κ_P and the water (in white), with high dielectric constant κ_W. The surface of separation between the two media is constructed as described in the text.

early realized by Cyrus Levinthal in 1969 [18]. Consequently, the internal energy of the system, which includes the water molecules and the protein, must be calculated a large number of times and the numerical complexity of the method chosen to account for the influence of water must be as low as allowed by the accuracy required to solve the problem.

Despite being regarded as one of the most accurate implementations of solvent influence, explicit water models are presently too computationally demanding, allowing only short simulations of peptides with a small number of residues to be performed. Another popular option is to use continuum models based on Poisson (PE) or Poisson–Boltzmann (PBE) equations [5,6,7,8], which are orders of magnitude faster than explicit solvent models, to account for the electrostatic part of the free energy of solvation [4,5]. Then, the nonelectrostatic part, which arises from the first layer of water molecules surrounding the solute and from the creation of the cavity, could be added in many ways, most of which are related to the *Solvent Accessible Surface Area* (SASA) [19].

However, it is worth stressing that only the total free energy of solvation is thermodinamically defined and experimentally measurable. Consequently, any partitioning of it is necessarily arbitrary and the free parameters contained in these continuum models (such as the dielectric constant κ_P of the protein,

the dielectric constant κ_W of the aqueous medium (see footnote 2), and the position of the surface that separates both regions (see fig. 3)), must be fit prior to use in order to agree with experiment or with more accurate methods.

In this section, the dielectric constants are set to their standard values, $\kappa_P = 1$ and $\kappa_W = 80$, and the characteristics of the surface of separation are modified. Rigorously speaking, one would need an infinite number of parameters to completely specify this surface. However, it is used herein a restricted subset of all the possible surfaces, namely, those that can be obtained by rolling a sphere of radius R_W on the outer side of the surface generated by adding R_+ to the Van der Waals radii [11] of each atom (see fig. 3). The volume that the rolling sphere does not intersect is considered to belong to the protein region. Typical values assigned to R_W and R_+ in the literature are [9,20]:

(1) $R_W = 0.0$ Å and $R_+ = 0.0$ Å, producing the *Van der Waals Surface*
(2) $R_W = 1.4$ Å and $R_+ = 0.0$ Å, producing the *Molecular Surface*
(3) $R_W = 0.0$ Å and $R_+ = 1.4$ Å, producing the *Solvent Accessible Surface*

These three surfaces are customarily used as the separation between the two dielectric media when the Poisson energy is calculated. However, it could be the case that a small change in the parameters R_W and R_+ significantly alters the properties of this particular energy landscape. In such a situation, the choice of the surface would be crucial to the behaviour of the system. Therefore, the robustness of the Poisson energy to changes in R_W and R_+ must be assessed.

To accomplish this, we study a particular system: the *de novo* designed protein known as Trp–Cage [21] (PDB code 1L2Y). The CHARMM molecular dynamics program [10,11] was used as a conformation generator. From the native conformation stored in the Protein Data Bank [22] a 10 ps heating dynamics [12], from $T = 0$ K to three different temperatures, $T = 500$ K, $T = 750$ K and $T = 1000$ K, was performed on the system. This was repeated 50 times for each final temperature with a different seed for the random numbers generator each time. The overall result of the process was the production of a set of 150 different conformations of the protein, 50 of which are *close to native*, 50 *partially unfolded* and 50 *completely unfolded* (see fig. 4). It is worth remarking that the short time in which the system was heated (10 ps) and the fact that there was no equilibration after this process cause the final temperatures

---

[11] As found in the `CHARMM23` [10,11] force field and implemented in the `pdb2pqr` utility included in the APBS program.

[12] The `c27b4` version of the CHARMM program was used. The molecular dynamics were performed using the *Leap Frog* algorithm therein implemented and the `param22` parameter set, which is optimized for proteins and nucleic acids. The water has been taken into account implicitly with the Dominy *et al.* [23] version of the Generalized Born Model built into the program.
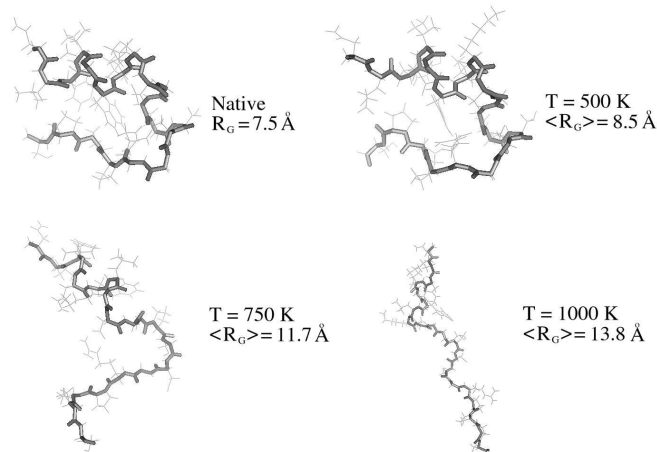
Fig. 4. Example conformations of the studied Trp–Cage protein. The native structure, taken from the Protein Data Bank is shown on the upper left corner. From left to right and from top to bottom, three particular conformations arbitrarily chosen from three different sets are depicted in order of decreasing similarity to the folded protein. The average radii of gyration in each set $\langle R_G \rangle$ and the one of the native structure are also presented.

(500, 750 and 1000 K) to be only *labels* for the three aforementioned sets of conformations. They are, by no means, the thermodynamical temperatures of any equilibrium state from which the structures are taken. This three sets of conformations are only meant to sample the representative regions of the phase space. In fig. 4, one arbitrarily chosen structure from each set is shown together with the native conformation. The average radius of gyration $\langle R_G \rangle$ of each set, depicted in the same figure, must be compared to the radius of gyration of the native state.

Using the finite differences APBS program [24], the Poisson–based electrostatic part of the solvation energy was numerically investigated in these conformations. To calculate this quantity, one must solve the Poisson equation twice. First, the energy of the system is computed assuming that a dielectric with $\kappa = \kappa_P$ fills the whole space. Second, one calculates the energy of the system with the dielectric geometry shown in fig. 3. Finally, the first quantity is substracted from the second to yield the solvation energy.

In order to test the reliability of the program and as an application of the first possible use, described in section 1, of the distance defined in this paper, the sensitivity of the Poisson energies to changes in the size of the grid L used to solve the differential equation was studied. For algorithmic reasons, the allowed values for L in APBS must be of the form $L_n = 32n + 1$, with n a positive integer. Consequently, the Poisson solvation energy of each of the 150 conformations of the protein was calculated [13] with $L = 33$, $L = 65$,

---

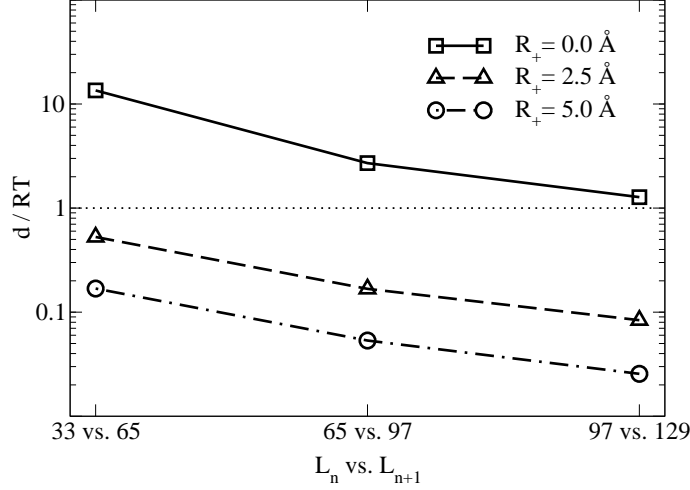[13] Boundary conditions flag `mdh` (non–interacting spheres with a point charges),

Fig. 5. Distance between instances of the Poisson solvation energy with different grid sizes. The y–axis corresponds to the distance d measured in units of RT (with $T = 300$ K) and the scale is logarithmic. Each point represents the comparison of the energies calculated with a smaller grid size L_n to those calculated with a larger one L_{n+1}. Results for different values of R_+ are shown. The value of R_W is set to 1.4 Å.

$L = 97$ and $L = 129$. All the measures were repeated for different values of R_+, (0.0, 2.5 and 5.0 Å) and R_W was fixed to 1.4 Å. Then, for each R_+, i.e. without changing the parameters, the distance (see eq. 8) between the energies calculated with a grid size L_n (playing the role of V_1) and the ones calculated with L_{n+1} (playing the role of V_2) was measured. The results are depicted in fig. 5.

Two conclusions may be drawn from these data. One one hand, as the size of the grid L increases, the distance d dimishes. This is consistent with the expectation that, when the accuracy of an approximation augments, the differences between an *exact* potential energy and its approximated counterpart tend to disappear. On the other hand, one sees that, for values of L between 97 and 129, the algorithm implemented in APBS has practically converged; in the sense that, for the worst case (namely, the one with $R_+ = 0.0$ Å), the distance between the energies calculated with $L = 97$ and $L = 129$ is of the order of the thermal noise.

Two remarks must be made about this last fact. First, the situation with $R_+ = 0.0$ Å being the worst is easily understood if one realizes that the discontinuity of the electric field in the surface of separation is larger if the latter is closer to the charges. Thus, a greater sensitivity to details is expected in this case. Second, it must be stressed that the distances depicted in fig. 5 place a lower bound in the distances to be considered meaningful when evaluating the

charges' grid mapping flag `spl2` (cubic B–spline discretization) and surface smoothing flag `smol` (simple harmonic averaging) were used in all the calculations.
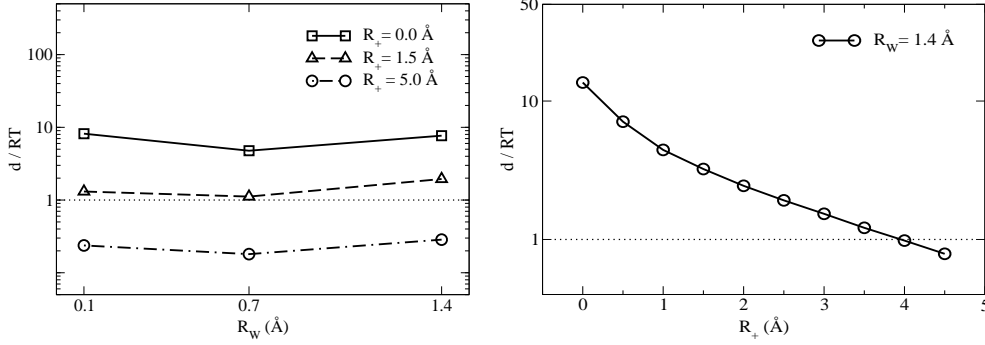
15

Fig. 6. Distance between instances of the Poisson solvation energy corresponding to different values of the free parameters. The y–axis corresponds to the distance d measured in units of RT (with $T = 300$ K) and the scale is logarithmic. Each point represents the comparison of the energies calculated with the value of the parameters in its x coordinate and the one calculated with the inmediately greater value (see text). The graphic on the left shows the results obtained when R_+ is held fixed and R_W is varied. When R_W is kept constant and R_+ is varied, the measured distance d is the one depicted in the graphic on the right.

robustness of the Poisson energy. For example, if the parameter R_W is slightly changed (keeping R_+ fixed to, say, 0.0 Å) any distance obtained, using a grid size of 97 or 129, that is below $\sim RT$ (see fig. 5) could not be associated to the lack of robustness of the Poisson solvation energy in that particular region of the parameter space, since it may be due to numerical inaccuracies of the algorithm.

Having this in mind, let us fix the grid size to 97 or 129 depending on the conformation [14] and evaluate the sensitivity of Poisson solvation energy to changes in the parameters R_W and R_+ that define the surface of separation, as an example of the application of the distance d to the second use proposed in section 1. This is done in the particular region of the parameter space which is typically explored in the literature: for R_W, the values 0.1, 0.7, 1.4 (the Van der Waals radius of a water molecule) and 2.8 Å; for R_+, the values from 0.0 to 5.0 Å in steps of 0.5 Å. When R_+ is kept constant and R_W is varied, the results on the left part of fig. 6 are obtained (only a few different values of R_+ are depicted). When, in turn, R_W is held fixed and R_+ is varied, one obtains the results shown on the right side of the same figure. In this second case, only the graphic corresponding to $R_W = 1.4$ Å is depicted, as the results for different values of R_W are practically identical. As in fig. 5, each point corresponds to the distance between the instances of the Poisson energy with the i–th value of the varying parameter and the one with the inmediately

---

[14] What was actually done was to choose $L = 97$ or $L = 129$ in order to keep the length of the grid cell below 0.5 Å in each dimension. In practice, this led to using $L = 97$ for the most compact and globular conformations and $L = 129$ for the most extended ones.

greater $(i+1)$–th value. Of course, if two instances with very distant values of the parameters are compared, the measured distance is much larger than the values depicted in fig. 6. However, this is not relevant to assess the robustness, since only the distance between instances corresponding to slightly different parameters must be small in order to render a fit significant.

From the data shown in fig. 6, one may extract some relevant conclusions. First, the two situations depicted are different in an important sense: while the robustness increases (d decreases) as one moves towards larger values of R_+ holding R_W constant, it does not change significantly in the opposite situation (i.e. increasing R_W with fixed R_+). The same behaviour may be inferred from the fact that, on the left side of fig. 6, graphics corresponding to different values of R_+ are found at different heigths, whereas, on the right side, the data with different values of R_W produce almost identical results (this is not shown for the sake of visual comfort). The second important fact that must be pointed out is that, in agreement with what one would expect, the robustness of the Poisson solvation energy is minimum when the surface of separation is placed close to the molecule (i.e., small values of R_+). In the left graphic of fig. 6, one sees that, when R_+ is of the size of the water molecule radius (1.4 Å), the distance between instances of the potential energy produced by a small change in R_W approximately reaches the largest numerical indetermination in fig. 5 and, consequently, the Poisson energy may be considered robust to such a change. In the right part of fig. 6, one finds that an equivalent level of robustness is only achieved at values of R_+ around 3.0 Å if R_W is held fixed and what is changed is R_+.

To summarize, one may conclude that the robustness of the Poisson–based electrostatic part of the solvation energy steadily increases when the surface that separates the two dielectric media is moved further away from the macro-molecular solute. The largest value of the distance d is of the order of $10RT$ when the surface of separation is placed on the Molecular or Van der Waals Surface ($R_+ = 0.0$ Å) and the sensitivity to parameter changes approximately reaches the numerical indetermination of the algorithm used when the surface is one layer of water molecules away from the protein.

4   Conclusions

When calculating a term or the totality of a potential energy function in complex systems, two situations are often faced: the necessity to evaluate the relative accuracy of an approximation or an algorithm respect to a more precise calculation and the need to assess the significance of a free parameters' fit. Herein, a distance between two different instances of the same potential energy function has been devised, which may be used to answer the two pre-

ceding questions by making meaningful statistical statements about the way in which energy differences are modified when changing the algorithm or the parameters.

In section 3, a practical example of the two cases is given by studying the sensitivity of the Poisson–based electrostatic part of the solvation energy to such changes. This example is useful, on one hand, to show that the distance behaves consistently in a real situation and, on the other hand, to estimate the robustness of the Poisson energy when small changes are performed on the ideal surface that separates the protein cavity from the aqueous media. It is shown that this robustness, both to changes in R_W and in R_+, increases as the surface is moved further away from the macromolecule, being $d \sim 10RT$ when the surface is placed at zero distance from the Van der Waals volume of the protein and reaching the numerical indetermination at a distance of around a layer of water molecules (~ 3.0 Å).

Acknowledgements

References

[1] K.A. Dill, Additivity principles in biochemistry, *J. Biol. Chem.* **272** (1997) 701–704.

[2] K.A. Dill, Polymer principles and protein folding, *Prot. Sci.* **8** (1999) 1166–1180.

[3] J.L. Alonso, G.A. Chass, I.G. Csizmadia, P. Echenique and A. Tarancón, Do theoretical physicists care about the protein folding problem?, in: R.F. Álvarez-Estrada *et al.*, eds., *Meeting on Fundamental Physics 'Alberto Galindo'* (Madrid, Aula Documental, 2004) (`arXiv:q-bio.BM/0407024`).

[4] T. Lazaridis and M. Karplus, Thermodynamics of protein folding: a microscopic view, *Biophys Chem.* **100** (2003) 367–395

[5] B. Roux, Implicit solvent models, in: O. Becker, A.D. MacKerrel, B. Roux and M. Watanabe, eds., *Computational Biophysics* (Marcel Dekker Inc., New York, 2001).

[6] B. Honig and A. Nicholls, Classical electrostatics in biology and chemistry, *Science* **213** (1995) 1144–1149.

[7] M. Orozco and F.J. Luque, Theoretical methods for the representation of solvent in biomolecular systems, *Chem. Rev.* **100** (2000) 4187–4225.

[8] C. Zhang, S.R. Kimura, Z. Weng, S. Vadja, R.C. Brower and C. Delisi, The waters of life, *J. Frank. Inst.* **335** (1997) 231–240.

[9] M.L. Connolly, Analytical molecular surface calculation, *Journal of Applied Crystallography* **16** (1983) 548–558.

[10] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan and M. Karplus, CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations, *J. Comp. Chem.* **4** (1983) 187–217,

[11] A.D. MacKerell Jr., B. Brooks, C.L. Brooks, III, L. Nilsson, B. Roux, Y. Won and M. Karplus, CHARMM: The Energy Function and Its Parameterization with an Overview of the Program, in: P.v.R. Schleyer et al., eds., *The Encyclopedia of Computational Chemistry* (John Wiley & Sons, Chichester, 1998) 271–277.

[12] W.H. Press, S.A. Teukolsky, W.T. Vetterling and B.P. Flannery, Numerical recipes in C. The art of scientific computing, Second ed., (Cambridge University Press, New York, 2002).

[13] P.R. Bevington and D.K. Robinson, Data reduction and error analysis for the physical sciences, Third ed., (Mc. Graw–Hill, New York, 2003).

[14] V.S. Pande, A. Grosberg and T. Tanaka, How accurate must potentials be for successful modeling of protein folding?, *J. Chem. Phys.* **103** (1995) 9482–9491.

[15] A.F. Pereira de Araújo and T.C. Pochapsky, Monte Carlo simulations of protein folding using inexact potentials: How accurate must parameters be in order to preserve essential features of the energy landscape?, *Folding & Design* **1** (1996) 299–314.

[16] A.F. Pereira de Araújo and T.C. Pochapsky, Estimates for the potential accuracy required in realistic folding simulations and structure recognition experiments, *Folding & Design* **2** (1997) 135–139.

[17] B. Hess, Stochastic concepts in molecular simulation, *Ph.D. Thesis* (Universal Press, University of Gronigen, Veenendaal, 2002) 12–14.

[18] C. Levinthal, How to fold graciously, in: J.T.P. DeBrunner and E. Munck, eds., *Mossbauer Spectroscopy in Biological Systems*, Proceedings of a meeting held at Allerton House (University of Illinois Press, Monticello, Illinois, 1969) 22–24.

See also `http://brian.ch.cam.ac.uk/~mark/levinthal/levinthal.html`.

[19] D. Eisenberg and A.D. McLachlan, Solvation energy in protein folding and binding, *Nature* **319** (1986) 199.

[20] B. Lee and F.M. Richards, The interpretation of protein structures: estimation of static accessibility, *J. Mol. Biol.* **55** (1971) 379–400.

[21] J.W. Neidigh, R.M. Fesenmeyer and N.H. Andersen, Designing a 20–residue protein, *Nat. Struct. Biol.* **9** (2002) 425.

[22] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov and P.E. Bourne, The Protein Data Bank, *Nucleic Acids Research* **28** (2000) 235–242.

[23] B. Dominy and C.L. Brooks III, Development of a Generalized Born Model Parameterization for Proteins and Nucleic Acids, *J. Phys. Chem.* **103** (1999) 3765–3773.

[24] N.A. Baker, D. Sept, S. Joseph, M.J. Holst and A.J. McCammon, Electrostatics of nanosystems: Application to microtubules and the ribosome, *Proc. Natl. Acad. Sci.* **18** (2001) 10037–10041.